Horizontal visibility graphs from integer sequences

Lucas Lacasa*
School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E14NS (UK)
(Dated: June 30, 2016)

The Horizontal Visibility Graph (HVG) is a graph-theoretical representation of a time series and builds a bridge between dynamical systems and graph theory. In recent years this representation has been used to describe and theoretically compare different types of dynamics and has been applied to characterize empirical signals, by extracting topological features from the associated HVGs which have shown to be informative on the class of dynamics. Among some other measures, it has been shown that the degree distribution of these graphs is a very informative feature that encapsulates nontrivial information of the series's generative dynamics. In particular, the HVG associated to a bi-infinite real-valued series of independent and identically distributed random variables is a universal exponential law $P(k) = (1/3)(2/3)^{k-2}$, independent of the series marginal distribution. Most of the current applications have however only addressed real-valued time series, as no exact results are known for the topological properties of HVGs associated to integer-valued series. In this paper we explore this latter situation and address univariate time series where each variable can only take a finite number n of consecutive integer values. We are able to construct an explicit formula for the parametric degree distribution $P_n(k)$, which we prove to converge to the continuous case for large n and deviates otherwise. A few applications are then considered.

I. INTRODUCTION

In recent years methods of network science have been applied to describe the structure of time series and signals, proposing mappings and transformations from series to graphs with the aim of making some sort of graph-theoretical time series analysis. Among other approaches [1–5], the family of visibility algorithms [6, 7] is a collection of recipes which map an ordered sequence of N numbers to a graph of N vertices where an edge between each two vertices exist if a certain geometric criterion is fulfilled in the sequence. Accordingly these methods have been shown to be very fruitful to give a topological characterization of time series and dynamics. In particular, it has been shown that (i) both the structure of complex, irregular time series and nontrivial ingredients of its underlying dynamics are inherited in the topology of the visibility graphs, and therefore (ii) simple topological properties of the graphs can be used as time series features for description and automatic classification purposes. Examples include a topological characterization of chaotic series (and routes to chaos) [12–14] or stochastic series [8, 20, 22, 28], and the method has been used for the description and classification of empirical time series appearing in physics [15–19, 23–25], physiology [26, 27], neuroscience [29] or finance [21, 30] to cite only a few examples.
In most of the practical applications, the time series under study are real-valued. As a matter of fact, the set of rigorous results that have appeared in recent years assume such thing. But how does the scenario changes when the time series under study can only take a finite set of integer values, or in other words, when the dynamics run over a finite field? From a combinatoric and number theoretic viewpoint (where integer sequences abound) this is an interesting question on itself. It is also relevant from a dynamical viewpoint, as in areas such as Markov Chain theory, symbolic dynamics or arithmetic dynamics integer-valued this is indeed the correct setting. Finally, while in practical applications empirical time series are assumed to be real-valued there are nonetheless circumstances where the empirical time series are inherently integer-valued.

In this work we partially fill this gap and propose as a first study to explore the properties of Horizontal Visbility Graphs (HVG) associated to random and uncorrelated integer-valued series. We focus on the degree distribution of these graphs as this is a metric which has been shown to be highly informative in the real-valued (continuous) case [32]. In this continuous case it was proved [7] that uncorrelated random processes have a universal exponential degree distribution, independent of the marginal distribution. Here we show that when the series takes values from $\{1, 2, \ldots, n\}$) the exponential law is only reached asymptotically (for $n \to \infty$), and for finite n the distribution deviates. The rest of the paper goes as follows: in section II we describe the method of HVG along with a few relevant properties and we outline the benchmark result obtained for real-valued series . In section III we make a describe our

*Electronic address: l.lacasa@qmul.ac.uk

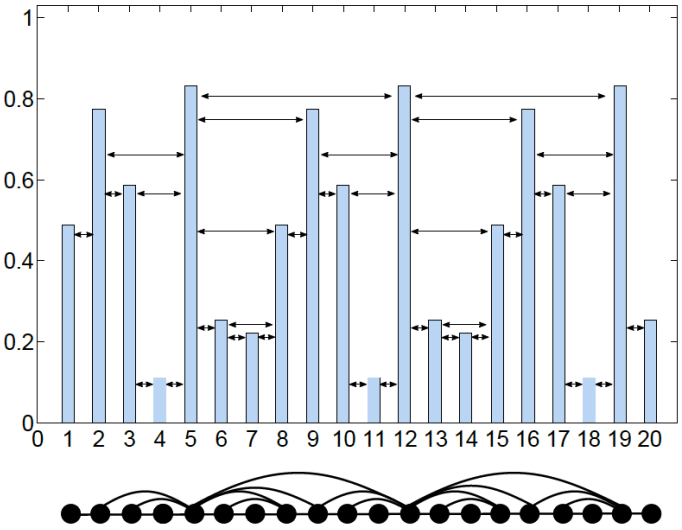

FIG. 1: Sample time series of 20 (real-valued) data and its associated horizontal visibility graph (HVG).

findings for the integer-valued case, which culminate with a closed expression for the parametric degree distribution. In section IV we conclude.

II. PRELIMINARIES

Let $\{x_1, \dots, x_N\}$, $x_i \in \mathbb{F}$ (finite or infinite field) be a sequence of N data. Its horizontal visibility graph HVG is defined as an undirected graph of N vertices, where each vertex is labelled in correspondence with the ordered datum x_i, so that x_1 is related to vertex $i = 1$, x_2 to vertex $i = 2$, and so on. Then, two vertices i, j (assume $i < j$ without loss of generality) share an edge if and only if $x_k < \inf(x_i, x_j)$, $\forall k : i < k < j$. This is an ordering criterion which can be visualized in figure 1.
HVG is a non crossing graph as described in algebraic combinatorics [9, 10] and can be proved to be invariant under monotonic transformations in the series [11]. It is therefore an order statistics of the associated process, which among other things means that the structure of this graph is not dependent upon the marginal distribution $f(x)$. In particular, the following result was found [7] for the degree distribution of the HVG associated to white noise:

Theorem 1. (Continuous case) *Consider a (bi-infinite) time series* $\{\dots, x_{-1}, x_0, x_1, \dots\}$ *of identically and independently distributed random variables extracted from a continuous marginal distribution* $f(x)$. *Then* $\forall f(x)$ *the associated HVG has a universal degree distribution* $P(k) = (1/3)(2/3)^{k-2}$

Our aim in this paper is to study the analogous statement in the case where instead of having $x \in \mathbb{R}$, the data take values over a finite field $\mathbb{F}$, which for simplicity we consider to be a subset of the integers. Let thus consider a (bi-infinite) time series $\{\dots, x_{-1}, x_0, x_1, \dots\}$ of identically and independently distributed random variables sampled from a uniform discrete distribution of n integers: $\forall t,\ x_t = \xi,\ \xi \sim \mathcal{U}\{1, 2, \dots, n\}$, and let us call $P_n(k)$ the degree distribution of the associated HVG. In figure 2 we plot in semi-log scales the numerical estimate of $P_n(k)$ for $n = 2, 3, 4, 8, 16, 32, 64, 128, 256, 512$ and 1024 (in every case we have generated a time series of 10^5 data). As n increases, we can see how $P_n(k)$ approaches the universal exponential shape found in the continuous case (red dashed line), this convergence being from above for $k = 2, 3$ and from below for $k \geq 4$. According to theorem 1, one should indeed expect $\lim_{n\to\infty} P_n(k) = P(k)$ as for large n the marginal distribution of the time series approaches a continuous form, however numerical evidence suggests that for small values of n the deviations from this law are large. In the rest of the paper we develop a combinatorial framework to explicitly compute $P_n(k)$ for finite n.

III. INTEGER-VALUED SERIES: THEORETICAL DERIVATION OF $P_n(k)$

We start by noting that $P_n(k)$ is equivalent to the probability that an arbitrary node of the graph (whose associated datum is for convenience denoted x_0) has degree k, that is x_0 has horizontal visibility of exactly k other data. Among these k neighbours, there always exist two *bounding* data (at right and left hand side respectively), as indeed the

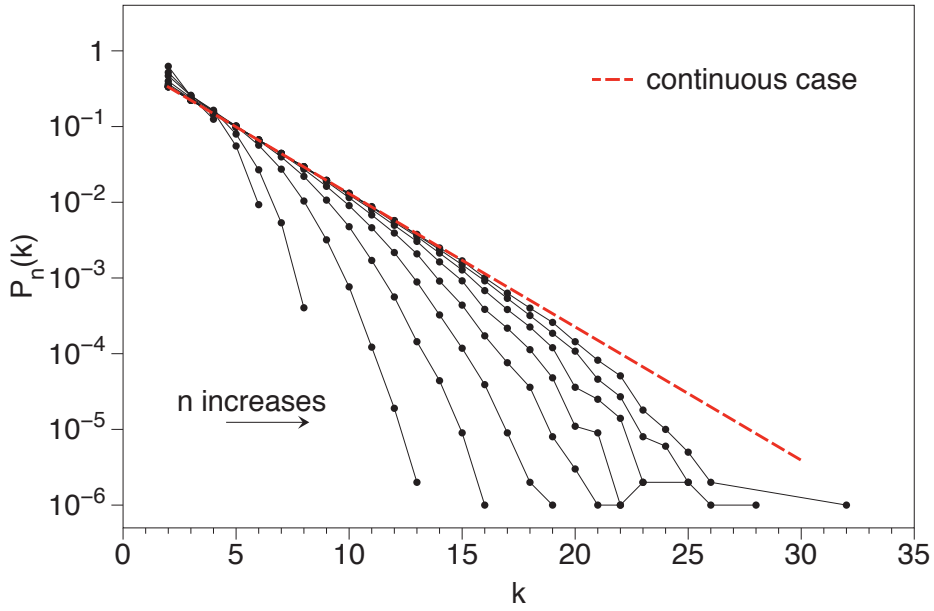

FIG. 2: Semilog plot of the numerical values of $P_n(k)$ for $n = 2, 3, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}$ extracted from series of $N = 10^5$ i.i.d. uncorrelated random variables $\xi \in \mathcal{U}\{1, \ldots, n\}$

minimum degree is $k = 2$. The remaining $k-2$ data are located among the bounding data. These *inner* data can be ordered (sorted by size) in only $k-1$ different ways, and it is easy to prove that one can label each of the $k-1$ configurations as C_i, $i = 0, 1, \ldots, k-2$ where the index i determines the number of inner data placed at the left hand side of x_0 (that is, the number of inner data taking place 'before' in the time series). In other words, C_i is the configuration for which out of the free $k-2$ visible inner data, i of them are placed before x_0 and $k-2-i$ are placed after x_0. On top of this, note that an arbitrary number of *hidden* data can take place after each inner datum. These hidden data don't contribute to the degree but play an important role in the computation of the degree probabilities. We therefore split $P_n(k)$ accordingly

$$P_n(k) = \sum_{i=0}^{k-2} P_{nk}[C_i], \tag{1}$$

So far, the construction follows the one elaborated in the continuous case. Note however that in the discrete case $n < \infty$, by construction not all C_i are admissible given a concrete value of n. For instance, it is easy to see that for $n = 2$, $P_{2k}(C_{i>1}) = 0$. Actually, it is easy to prove that given n, the largest admissible degree $k_{\max}(n) = 2n$, and thus $P_n(k > 2n) = 0$. These restrictions were not present in the continuous case and it can be proved that in the general case we have the following:

Lemma 1. *Given n and k, after only counting admissible configurations eq. 1 is effectively reduced into*

$$P_n(k) = \sum_{i=\max(k-n-1,0)}^{\min(n-1,k-2)} P_{nk}[C_i]$$

Proof. The proof is based on counting the minimal number of distinct symbols in the distribution for a given configuration to be feasible. To be able to allocate i inner data and a bounding data at the left hand side of x_0, one needs for those to be visible that $n \geq i+1$, therefore $i \leq n-1$ (this proofs the upper limit). Respectively for the right hand side one needs $n \geq k-1-i$, therefore $i \geq k-n-1$. □

Notice that the discrete nature of each random variable x precludes using integrals to calculate probabilities, so the approach followed to prove theorem 1 is not valid here anymore. We split again the computation by conditioning each configuration to the value of x_0. Applying Bayes theorem, we have

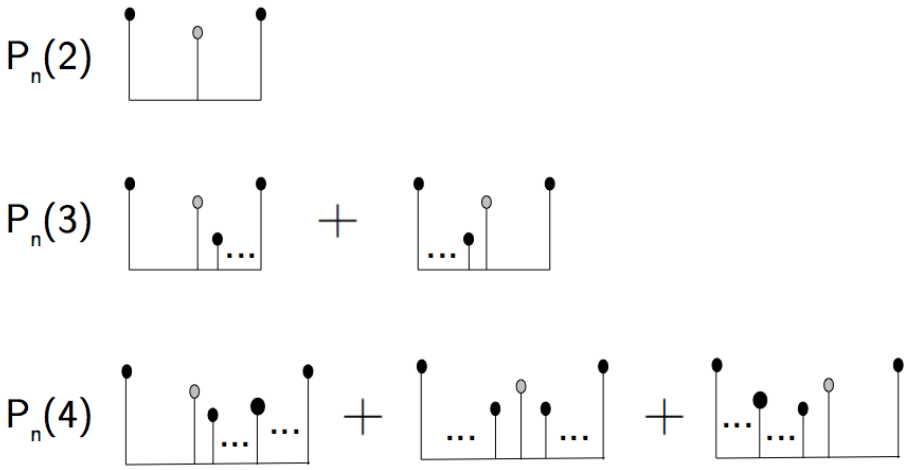

FIG. 3: Sample diagrams for $P_n(2)$, $P_n(3)$ and $P_n(4)$.

$$P_{nk}[C_i] = \sum_{m=1}^{n} P_{nk}[C_i|x_0 = m]S_n(x_0 = m) \tag{2}$$

where $P_{nk}[C_i|x_0 = m]$ is the probability of C_i taking place, conditioned to x_0 taking a particular value $x_0 = m$ (where $m \in [1, n]$) and $S_n(x_0 = m)$ is the probability that x_0 takes indeed the value m. As we assume uniformly distributed random variables we have $\forall m S_n(x_0 = m) = 1/n$ but this condition can be removed if needed. We define $\mathscr{P}_{nki}(m) := P_{nk}[C_i|x_0 = m]S_n(x_0 = m)$. Now, again the fact that n is finite necessarily forbids some events, effectively reducing the number of terms in equation 2. This is summarized in the following proposition.

Lemma 2. *Given n and k, after only counting admissible events eq. 2 is effectively reduced into*

$$P_{nk}[C_i] = \sum_{m=\max(i+1,k-1-i)}^{n} \mathscr{P}_{nki}(m)$$

Proof. To be able to allocate i inner data and a bounding data at the left hand side of $x_0 = m$, one needs for those to be visible that $m \geq i + 1$. Respectively for the right hand side one needs $m \geq k - 1 - i$. □

Once we have formally splitted the computation of $P_n(k)$ into configurations and conditioned to different values of x_0 we are ready to derive rigorously this parametrized distribution. Before attempting to find a general expression for $P_n(k)$ for illustrative purposes we need to study some particular cases first. All over these examples we make use of lemmas 1 and 2.

A. Illustrative examples: $P_2(2)$, $P_n(2)$, $P_3(3)$, $P_n(3)$ and $P_4(4)$

$P_2(2)$. For $n = 2$, the time series only takes values from $\{1, 2\}$, and $S_2(m) = 1/2$. The probability for each degree can be easily computed in explicit form. There is only one configuration with neither *inner* nor *hidden* data, just the *seed* and two *bounding* data (see figure 3). We can condition x_0 to be both 1 and 2, so trivially

$$P_2(2) = \mathscr{P}_{220}(1) + \mathscr{P}_{220}(2) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 5/8,$$

where each contribution follows the structure of the diagram $\mathscr{P}_{220}(\cdot) \equiv [B][S][B]$, where $[B]$ and $[S]$ denote the probability of the bounding and seed data respectively. Note that when $m = 1$ the bounding data will indeed 'bound' x_0 regardless of their value.

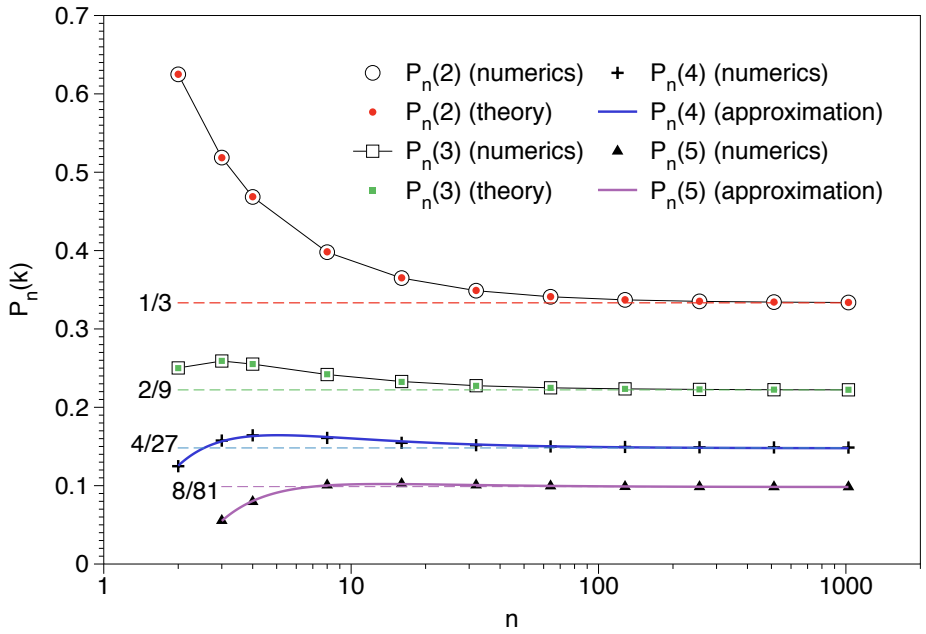

FIG. 4: log-linear plot of the numerical values of $P_n(k)$ for $k < 6$ and $n = 2, 3, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}$ extracted from series of $N = 10^5$ i.i.d. uncorrelated random variables $\xi \in \mathcal{U}\{1, \ldots, n\}$. The dashed lines correspond to the values for the continuous case and solid points are the predictions of the theory. The solid lines for the cases $k = 4, 5$ are fittings to rational approximations $P_n(k) \approx \alpha + \beta/n + \gamma/n^2$

$P_n(2)$. The result for $P_2(2)$ can be readily generalized for an arbitrary n. In this case, $S(m) = 1/n$ and we can easily see that the probability of a bounding data is not simply $1/n$ but will depend on the conditioning of x_0, in such a way that for $x_0 = m$ one has

$$[B] \to [B(m)] = \frac{n+1-m}{n}$$

It is therefore easy to prove by induction that

$$P_n(2) = \sum_{m=1}^{n} [B(m)][S_n(m)][B] = \frac{1}{n^3} \sum_{m=1}^{n} (n+1-m)^2 = \frac{2n^2+3n+1}{6n^2} \tag{3}$$

Note that $\lim_{n\to\infty} P_n(2) = 1/3$ and thus the discrete case converges to the continuous case asymptotically (see figure 4 for a comparison with numerical values).

$P_3(3)$. According to the previous formula we find $P_3(2) = 14/27$. Now $P_3(3)$ is the result of two configurations C_0 and C_1 which are actually symmetric, therefore $P_3(3) = P_{33}(C_0) + P_{33}(C_1) = 2P_{33}(C_0)$ so we focus on C_0 without loss of generality. As $k = 3$, in C_0 we have an inner datum (and associated hidden structure) at the right hand side of x_0 (see figure 3) and by virtue of lemma 2 $m \geq 2$, thus

$$P_{33}(C_0) = \mathscr{P}_{330}(2) + \mathscr{P}_{330}(3)$$

By construction,

$$\mathscr{P}_{330}(2) = [B(2)][S(2)][IH(2)][B(2)] = \frac{2}{3} \cdot \frac{1}{3} \cdot \left[\frac{1}{3} \cdot \sum_{k=0}^{\infty} \left(\frac{1}{3}\right)^k\right] \cdot \frac{2}{3} = \frac{4}{54}$$

where: (i) the bounding datum can be either 2 or 3, thus has probability 2/3, (ii) we have put together the inner datum (which could only be equal to 1) and its hidden structure (an arbitrary number of data hidden by the inner datum). As this inner datum needs to be equal to 1, then the hidden data can only take the value 1, thus contributes with a geometric series with common ratio 1/3. Conversely, for $m = 3$ the bounding data can only take one value,

the inner datum is free to take the values 1 or 2, and for this latter case its hidden structure can be formed by 1s and 2s. Accordingly we find

$$\mathscr{P}_{330}(3) = [B(3)][S(3)][IH(3)][B(3)] = \frac{1}{3}\cdot\frac{1}{3}\cdot\left[\frac{1}{3}\cdot\sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^k + \frac{1}{3}\cdot\sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k\right]\cdot\frac{1}{3} = \frac{3}{54}$$

Altogether, $P_3(3) = 14/54$.

$\underline{P_n(3)}$ We can again generalize the latter result for an arbitrary n, as $P_n(3) = 2P_{n3}(C_0)$, where for an arbitrary n according to lemma 2 we have $P_{n3}(C_0) = \sum_{m=2}^{n}\mathscr{P}_{nki}(m)$ and $\mathscr{P}_{nki}(m) = [B(m)][S_n(m)][IH(n,m)][B(m)]$, where it is easy to prove by induction that

$$[IH(n,m)] = S_n(m)\sum_{p=1}^{m-1}\left[\sum_{k=0}^{\infty}\left(\frac{p}{n}\right)^k\right] = \sum_{p=1}^{m-1}\frac{1}{n-p}$$

Now this last expression can be summed up in terms of harmonic numbers. Altogether,

$$P_n(3) = \frac{2}{n}\sum_{m=2}^{n}\left(\frac{n+1-m}{n}\right)^2\sum_{p=1}^{m-1}\frac{1}{n-p} \approx \frac{4n^2+3n-1}{18n^2}, \tag{4}$$

where the last approximation is valid for $n > 4$ (for $n = 2, 3, 4$ $P_n(3)$ take the values $1/4$, $14/54$ and $49/192$ respectively). Comparison with numerics is reported in figure 4. Note that again we have $\lim_{n\to\infty} P_n(3) = 2/9 = P(3)$, and therefore the discrete case again converges to the continuous case for large n.

From the last particular case $P_3(3)$ we have also learned that the the probability contribution of an inner-hidden data structure indeed also depends on the conditioning of x_0, as a summation dependent on m emerges due to the fact that both the inner and the hidden data are allowed to take different values depending on the conditioning on x_0. As this is in relation to both n and the number of inner variables, then it is sensible to write formally $[IH] \equiv [IH(m,n,i)]$, i.e. the probability of this structure is a function of the conditioning of x_0, n and the configuration C_i. The concrete dependence will be evident only after the next particular case, but for now we can give a formal expression for a general $\mathscr{P}_{nki}(m)$ as

$$\mathscr{P}_{nki}(m) = [B(m)][IH(m,n,i)][S_n(m)][IH(m,n,k-2-i)][B(m)] \tag{5}$$

$\underline{P_n(4)}$. This is the next nontrivial case. By symmetry we have $P_n(4) = 2P_{n4}(C_0) + P_{n4}(C_1)$, but now C_0 and C_1 are qualitatively different configurations. Whereas for C_1 there is exactly one inner datum at each side of the seed, in C_0 we will find two *concatenated* inner data at the right hand side of the seed. In what follows we will see that is has a dramatic effect, as $IH(m,n,2) \neq IH(m,n,1)^2$. Let us consider the easier case C_1 first. Formally, each side of the seed is independent and therefore

$$P_{n4}[C_1] = \sum_{m=2}^{n}[S_n(m)]\Big([IH(m,n,1)][B(m)]\Big)^2$$

where

$$[IH(m,n,1)] \equiv [IH(m,n)] = S_n(m)\sum_{p=1}^{m-1}\left[\sum_{k=0}^{\infty}\left(\frac{p}{n}\right)^k\right] = \sum_{p=1}^{m-1}\frac{1}{n-p}$$

Therefore

$$P_{n4}[C_1] = \sum_{m=2}^{n}\frac{(n+1-m)^2}{n^3}\left[\sum_{p=1}^{m-1}\frac{1}{n-p}\right]^2$$

On the other hand, for C_0 we have

$$P_{n4}[C_0] = \sum_{m=3}^{n}[S_n(m)][IH(m,n,0)][IH(m,n,2)][B(m)]^2,$$

where $[IH(m,n,0)]=1$ and m starts at $m=3$ according to lemma 2. The key ingredient of this configuration is of course $[IH(m,n,2)]$. To understand its structure, let us consider $n=4$ for simplicity. Then after a bit of algebra

$$IH(3,4,2)=\left[\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{1}{4}\right)^k\right]\cdot\left[\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{2}{4}\right)^k\right],$$

$$IH(4,4,2)=\left[\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{1}{4}\right)^k\cdot\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{2}{4}\right)^k+\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{1}{4}\right)^k\cdot\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{3}{4}\right)^k+\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{2}{4}\right)^k\cdot\frac{1}{4}\sum_{k=0}^{\infty}\left(\frac{3}{4}\right)^k\right].$$

The general structure of $IH(m,n,i)$ can be proved by induction. Intuitively, it is a sum of terms where each term combines the product of i contributions where in each case the hidden variables can take a different number of possible values. Essentially, we are enumerating the different possible arrangements of i inner data (and an arbitrary number of hidden data among each inner datum), where the inner data take values from $\{1,\ldots,m-1\}$ ($m-1$ is the upper bound as one needs to leave room for the bounding datum). With the extra condition that the seed takes the value m and that all inner data are visible by construction, for i inner data there are $\binom{m-1}{i}$ different ways of giving values to the inner data.

B. An exact formula for $P_n(k)$

Let

$$f(z)=\frac{1}{n}\sum_{k=0}^{\infty}\left(\frac{z}{n}\right)^k=\frac{1}{n-z} \tag{6}$$

Then one can show that

$$[IH(m,n,i>0)]=\sum_{j_1=1}^{m-1}\sum_{j_2=j_1+1}^{m-2}\cdots\sum_{j_i=j_{i-1}+1}^{m-i}f(j_1)\cdot f(j_2)\cdots f(j_i)=\prod_{l=1}^{i}\sum_{j_l=j_{l-1}+1}^{m-i+l-1}f(j_l), \tag{7}$$

where we need to define $j_0:=0$ and $[IH(m,n,i=0)]:=1$. The general solution to the parametric degree distribution $P_n(k)$ is then provided by lemmas 1 and 2 together with eqs. 5,6 and 7. We are thus ready to put this altogether:

$$P_n(k)=\sum_{i=\max(k-n-1,0)}^{\min(n-1,k-2)}\sum_{m=\max(i+1,k-1-i)}^{n}\frac{n+1-m}{n^3}\left[\prod_{l=1}^{i}\sum_{j_l=j_{l-1}+1}^{m-i+l-1}f(j_l)\right]\left[\prod_{l=1}^{k-2-i}\sum_{j_l=j_{l-1}+1}^{m-i+l-1}f(j_l)\right], \tag{8}$$

where $j_0:=0$. This formula gives a recipe to compute $P_n(k)$ for arbitrary values of n and k. Unfortunately we have not been able to find an algebraic enumeration and an associated generic algebraic closed form for eq. 8. On the other hand, for a fixed k we have seen that one finds suitable rational functions such as eqs. 3 or 4. We conjecture

$$f(n):=P_n(k)|_{k\text{ fixed}}=\frac{An^2+Bn+C}{Dn^2}\equiv\alpha+\frac{\beta}{n}+\frac{\gamma}{n^2},$$

where $A,B,C,D\in\mathbb{N}^+$, $\alpha=A/D=(1/3)(2/3)^{k-2}$, $\beta=B/D$, $\gamma=C/D$. In figure 4 we plot this approximation (which is exact for $k=2,3$) for $k<6$, showing a perfect agreement with the numerics. We also find in that figure that for each k, the convergence to the continuous case $P(k)$ is faster as k increases.

C. Asymptotic approximation

An elementary asymptotic approximation for $P_n(k)$ can be found using a simple combinatorial argument. For a fixed n, $P_n(2)$ can be understood as the probability that an arbitrary datum is bounded, so $1-P_n(2)$ is the probability that a given datum is not bounded by its first neighbours. In the same line, for a fixed n, $P_n(k)$ is the probability

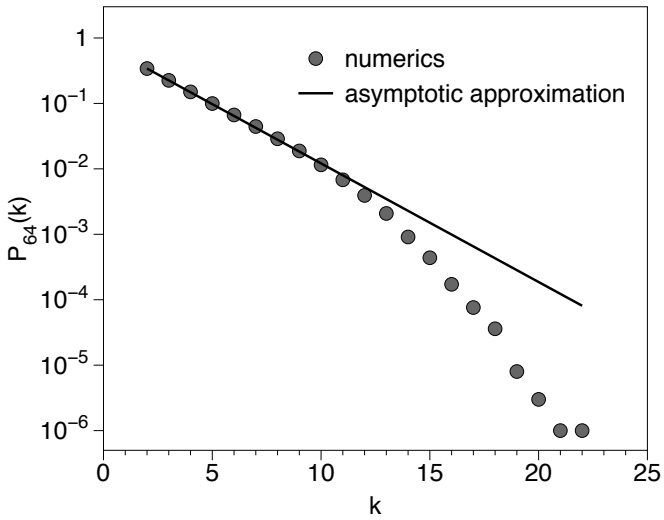

FIG. 5: log-linear plot of the numerical values of $P_{64}(k)$ extracted from series of $N = 10^5$ i.i.d. uncorrelated random variables $\xi \in \mathcal{U}\{1, \ldots, n\}$. The solid line corresponds to the asymptotic approximation (eq. 9 for $n = 64$), which works well for the exponentially decaying part of the distribuion.

that an arbitrary datum has at least visibility with $k-2$ inner data -whose probability can be approximated to $(1 - P_n(2))^{k-2}$- which is then bounded. This sort of 'Markovian' approximation gives

$$P_n^{\text{app}}(k) = P_n(2)(1 - P_n(2))^{k-2} = \frac{(2n^2 + 3n + 1)[1 - (2n^2 + 3n + 1)/6n^2]^{k-2}}{6n^2} \tag{9}$$

where the last step involves taking the limit of large n. This is an algebraic closed form equation, however this formula is not exact as it is not taking into account that inner data are correlated (i.e. the values that each inner data can take depend on the position of the inner data). Still, this approximation improves when n increases, and as the argument holds exactly in the limit $n \to \infty$ we have $\lim_{n\to\infty} P_n^{\text{app}}(k) = \lim_{n\to\infty} P_n(k)$, hence $P_n^{\text{app}}(k)$ and $P_n(k)$ are asymptotic. On the other hand, taking the limit in eq.9 we also have

$$\lim_{n\to\infty} P_n^{\text{app}}(k) = \frac{1}{3}\left(\frac{2}{3}\right)^{k-2},$$

concluding that $P_n(k)$ indeed converges to $P(k)$. It is easy to see that $P_n^{\text{app}}(k)$ is essentially an exponential approximation and, according to figure 2, gives good estimates of $P_n(k)$ for those values that comply to an exponential shape. In other words, the super-exponential cutoff that develops for large k for each n is badly approximated by $P_n^{\text{app}}(k)$, however for the range of values of k for which each distribution approaches an exponential decay, $P_n^{\text{app}}(k)$ should give a good match. A confirmation of this is shown for a particular example where $n = 2^6$ in figure 5.

IV. CONCLUDING REMARKS AND DISCUSSION

In recent years several rigorous results have been advanced within the theory of horizontal visibility graphs. Yet in all the cases the series under study was assumed to be real-valued. In this work we depart from this assumption and study the properties of the degree distribution $P_n(k)$ associated to a random uncorrelated series $\{x_1, x_2, \ldots\}$ that only takes a finite number of values $\mathbb{F} \subset \mathbb{Z}$. Note at this point that $\mathbb{F}$ does not need to be in the form $(1, 2, \ldots, n)$: as the HVG is invariant under monotonic transformations in the series, it is only required that $\mathbb{F} = (b_1, b_2, \ldots, b_n)$, where $b_i = b_{i-1} + c$ and $c \in \mathbb{N}^+$.

We have observed that for any finite n, $P_n(k)$ deviates from the universal shape $P(k)$ obtained in the continuous case, and confirmed analytically that $\lim_{n\to\infty} P_n(k) = P(k)$. As we have seen, moving from infinite to finite fields makes the problem considerably more difficult and involved to address analytically, however we have been able to show in an explicit way how to analytically compute $P_n(k)$ for an arbitrary n and k, although unfortunately we haven't found a closed algebraic form.

While the theory has been derived assuming (bi)infinite size series, we also found good convergence properties for finite

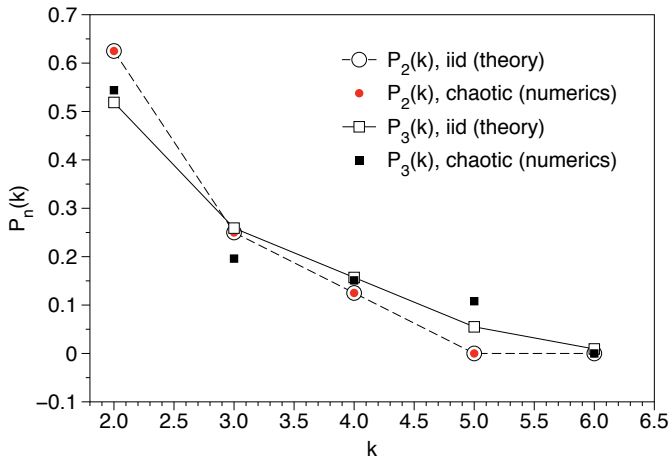

FIG. 6: $P_n(k)$ extracted from series of (empty symbols) the theory obtained for i.i.d. integer random variables $\xi \in \mathcal{U}\{1,\ldots,n\}$ and (solid symbols) a chaotic trajectory of 10^6 data points from $x_{t+1} = 4x_t(1-x_t)$, after coarse-graining into a symbolic sequence with n symbols via homogeneous partition of the phase space. Circles correspond to the case $n = 2$ while squares correspond to $n = 3$. We observe that $P_2(k)$ fails to discriminate both processes, $P_3(k)$ is already clearly different.

sizes. As an illustration, we consider the ability of this method to distinguish between a purely random, uncorrelated process (iid) and a deterministic, chaotic process generated by a fully chaotic logistic map $x_{t+1} = 4x_t(1-x_t)$, $x \in [0,1]$. The power spectrum of both processes is flat (and therefore both processes have delta-distributed autocorrelation functions), so this discrimination is nontrivial a priori. As a matter of fact, it is well known that HVG easily distinguishes both processes as their associated degree distribution is clearly different [7, 8]. In other words, in the limit $n \to \infty$, $P_n(k)$ easily discriminates these processes. What happens if we compare symbolic representations of both processes (namely, finite n)? To explore this, we proceed to construct an homogeneous partition of the interval $[0,1]$ into n non-overlapping cells of equal size, and accordingly we construct the integer series associated to a chaotic trajectory $\{x_t\}$, and we compare this series with a sequence of 'unbiased coin tosses' where the coin has n faces ($\xi \in \{1,2,\ldots,n\}$. In figure 6 we compare the theoretical shape of $P_n(k)$ associated to the unbiased iid process with the numerics obtained in the chaotic case, finding that while $P_2(k)$ seems to be identical in both processes, already $P_3(k)$ shows substantially deviations.

Incidentally, in practice one can always construct statistical tests to investigate the compliance to $P_n(k)$ for experimental sequences of finite size N (such as the case for $n = 2$ above). For instance, a Pearson χ^2 statistic can be used

$$\chi^2 = N \sum_{k=2}^{k_{\max}(n)=2n} \frac{[f_n(k) - P_n(k)]^2}{P_n(k)}, \tag{10}$$

where $f_n(k)$ is the observed (estimated) frequency and $P_n(k)$ is the theoretical frequency. χ^2 is the Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution with $2n-1$ degrees of freedom. One could build an hypothesis test where the null hypothesis is that the time series is uncorrelated, and apply this in a variety of empirical series and situations, such as to explore the normality conjecture of numbers such as π, $\sqrt{2}$ etc. In the example considered above it is not really necessary to use an hypothesis test for $n = 3$ as there is a clear deviation. For $n = 2$ we can apply it: the theoretical values are $P_2(2) = 5/8$, $P_2(3) = 1/4$, $P_2(4) = 1/8$ and for $k > 4$ $P_2(k) = 0$, so for a trajectory of $N = 10^6$ the statistic gives $\chi^2 \approx 0.413$ thus we cannot reject the null hypothesis (i.e. the method cannot discriminate for $n = 2$) as this value is much smaller than the critical ones.
The application of HVG to inherently discrete (symbolic) sequences in areas such as text analysis in linguistics or DNA sequencing in bioinformatices to cite just a couple are potential avenues for future research.

[1] Zhang J, Small M, Complex network from pseudoperiodic time series: topology versus dynamics. *Phys. Rev. Lett.* **96**, 238701 (2006).

[2] Kyriakopoulos F, Thurner S, Directed network representations of discrete dynamical maps, in *Lecture Notes in Computer Science* **4488**, 625–632 (2007)

[3] Xu X, Zhang J, Small M Superfamily phenomena and motifs of networks induced from time series. *Proc. Natl. Acad. Sci. USA* **105**, 19601-19605 (2008).

[4] Donner R V, Zou Y, Donges J F, Marwan N, Kurths J Recurrence networks: a novel paradigm for nonlinear time series analysis. *New J. Phys.* **12**, 033025 (2010).

[5] Donner R V, et al. The Geometry of Chaotic Dynamics - A Complex Network Perspective. *Eur. Phys. J. B* **84**, 653-672 (2011).

[6] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J.C. Nuño, From time series to complex networks: the visibility graph, *Proc. Nat. Acad. Sci. USA* **105(13)** (2008) 4972-4975.

[7] B. Luque, L. Lacasa, F. Ballesteros, J. Luque, Horizontal visibility graphs: Exact results for random time series, *Phys. Revi. E* **80(4)** (2009) 046103.

[8] L. Lacasa, On the degree distribution of horizontal visibility graphs associated to Markov processes and dynamical systems: diagrammatic and variational approaches, *Nonlinearity* **27** (2014) 2063-2093.

[9] S. Severini, G. Gutin, T. Mansour, A characterization of horizontal visibility graphs and combinatorics on words, *Physica A* **390**, 12 (2011) 2421-2428.

[10] P. Flajolet and M. Noy, Analytic combinatorics of non-crossing configurations, *Discrete Math.* **204** (1999) 203-229.

[11] L. Lacasa, R. Flanagan, Time reversibility from visibility graphs of nonstationary processes, *Phys. Rev. E* 92, 022817 (2015).

[12] B. Luque, L. Lacasa, F. Ballesteros, A. Robledo, Analytical properties of horizontal visibility graphs in the Feigenbaum scenario, *Chaos* **22**, 1 (2012) 013109.

[13] B. Luque, A. Núñez, F. Ballesteros, A. Robledo, Quasiperiodic Graphs: Structural Design, Scaling and Entropic Properties, *Journal of Nonlinear Science* **23**, 2, (2012) 335-342.

[14] A.M. Núñez, B. Luque, L. Lacasa, J.P. Gómez, A. Robledo, Horizontal Visibility graphs generated by type-I intermittency, *Phys. Rev. E*, **87** (2013) 052801.

[15] A. Aragoneses, L. Carpi, N. Tarasov, D.V. Churkin, M.C. Torrent, C. Masoller, and S.K. Turitsyn, Unveiling Temporal Correlations Characteristic of a Phase Transition in the Output Intensity of a Fiber Laser, *Phys. Rev. Lett.* 116, 033902 (2016).

[16] M. Murugesana and R.I. Sujitha1, Combustion noise is scale-free: transition from scale-free to order at the onset of thermoacoustic instability, *J. Fluid Mech.* 772 (2015).

[17] A. Charakopoulos, T.E. Karakasidis, P.N. Papanicolaou and A. Liakopoulos, The application of complex network time series analysis in turbulent heated jets, *Chaos* 24, 024408 (2014).

[18] P. Manshour, M.R. Rahimi Tabar and J. Peinche, Fully developed turbulence in the view of horizontal visibility graphs, *J. Stat. Mech.* (2015) P08031.

[19] C Liu, WX Zhou, WK Yuan, Statistical properties of visibility graph of energy dissipation rates in three-dimensional fully developed turbulence, *Physica A* 389, 13 (2010).

[20] WJ Xie, WX Zhou, Horizontal visibility graphs transformed from fractional Brownian motions: Topological properties versus the Hurst index. *Physica A* 390, 3592-3601 (2011).

[21] MC Qian, ZQ Jiang, WX Zhou, Universal and nonuniversal allometric scaling behaviors in the visibility graphs of world stock market indices. *Journal of Physics A* 43, 335002 (2010).

[22] XH Ni, ZQ Jiang, WX Zhou, Degree distributions of the visibility graphs mapped from fractional Brownian motions and multifractal random walks. *Physics Letters A* 373, 3822-3826 (2009).

[23] RV Donner, JF Donges, Visibility graph analysis of geophysical time series: Potentials and possible pitfalls, *Acta Geophysica* 60, 3 (2012).

[24] V. Suyal, A. Prasad, H.P. Singh, Visibility-Graph Analysis of the Solar Wind Velocity, *Solar Physics* **289**, 379-389 (2014)

[25] Y. Zou, R.V. Donner, N. Marwan, M. Small, and J. Kurths, Long-term changes in the north-south asymmetry of solar activity: a nonlinear dynamics characterization using visibility graphs, *Nonlin. Processes Geophys.* **21**, 1113-1126 (2014).

[26] J.F. Donges, R.V. Donner and J. Kurths, Testing time series irreversibility using complex network methods, *EPL* 102, 10004 (2013).

[27] S. Jiang, C. Bian, X. Ning and Q.D.Y. Ma, Visibility graph analysis on heartbeat dynamics of meditation training, *Appl. Phys. Lett.* **102** 253702 (2013).

[28] L. Lacasa, B. Luque, J. Luque and J.C. Nuño, The Visibility Graph: a new method for estimating the Hurst exponent of fractional Brownian motion, *EPL* 86 (2009) 30001.

[29] M Ahmadlou, H Adeli, A Adeli, New diagnostic EEG markers of the Alzheimer's disease using visibility graph, *J. of Neural Transm.* 117, 9 (2010).

[30] R. Flanagan and L. Lacasa, Irreversibility of financial time series: a graph-theoretical approach, *Physics Letters A* 380, 1689-1697 (2016)

[31] M. Newman, The structure and function of complex networks, *SIAM Review* 45, 167-256 (2003).

[32] B. Luque, L. Lacasa, Canonical horizontal visibility graphs are uniquely determined by their degree sequence (under review).